Dipl.-Informatiker Martin Eul

Klassifikationsserver – Standardklassifikationen im maschinenlesbaren Format

Die freie Verfügbarkeit öffentlicher Daten hat in den letzten Jahren unter dem Begriff "Open Data" immer mehr an Bedeutung gewonnen. Dabei ist nicht nur die Bereitstellung der Daten selbst relevant, sondern insbesondere die Möglichkeit, diese Daten in eigenen Anwendungen nutzen und weiterverarbeiten zu können – dass sie also in einer maschinenlesbaren Form zur Verfügung gestellt werden.

Neben den veröffentlichten statistischen Daten sind auch die Metadaten von essenzieller Bedeutung. Nur in Verbindung mit den Metadaten lassen sich statistische Werte sinnvoll interpretieren. In der Welt der Statistik spielen dabei Klassifikationen eine besondere Rolle. Sie ermöglichen eine Verdichtung der erhobenen Daten, da sie die zu beobachtenden Tatbestände in verbindliche Kategorien einteilen. Um die Vergleichbarkeit der statistischen Ergebnisse zu gewährleisten, berücksichtigen nationale Klassifikationen dabei europäische beziehungsweise internationale Klassifikationen.

Mit dem Klassifikationsserver der Statistischen Ämter des Bundes und der Länder (www.klassifikationsserver.de) steht ein System zur Recherche in nationalen Standardklassifikationen zur Verfügung. Die enthaltenen Informationen können darüber hinaus in verschiedenen Dateiformaten heruntergeladen oder von Anwendungen über eine Webservice-Schnittstelle abgefragt werden. Innerhalb der amtlichen Statistik in Deutschland leistet der Klassifikationsserver einen wertvollen Beitrag zur Standardisierung der Produktionsprozesse.

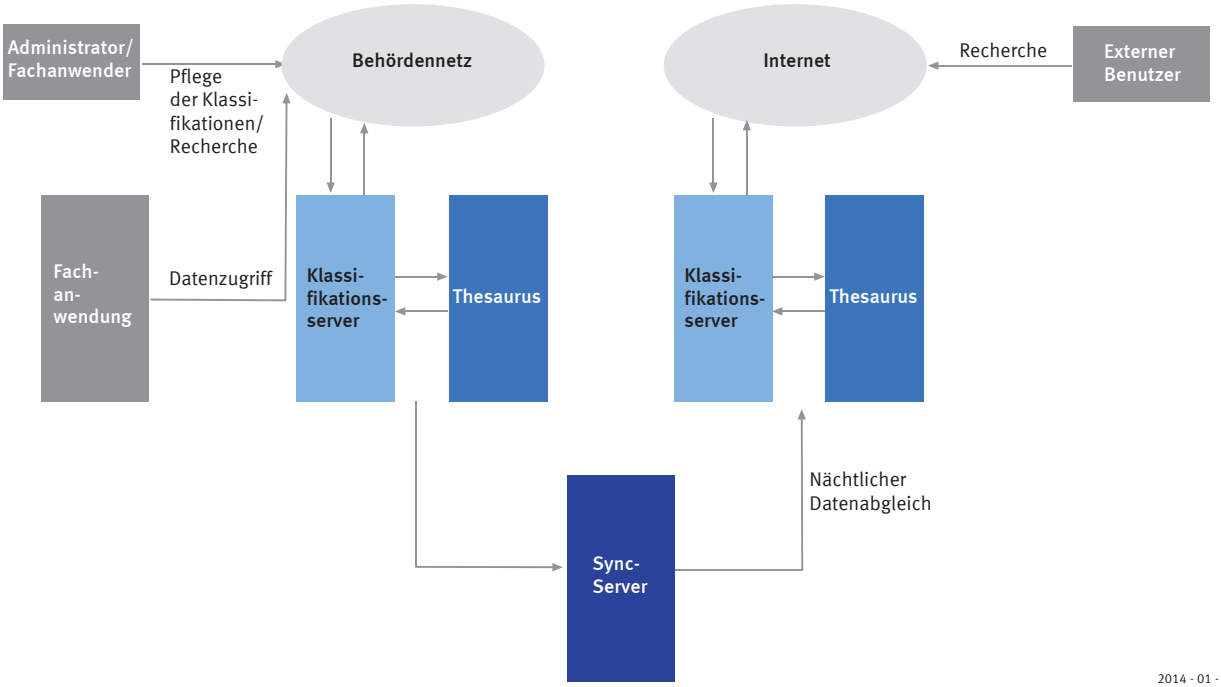
Projektziele

Die Statistischen Ämter des Bundes und der Länder verfügen bereits seit einigen Jahren über eine Webanwendung, die auch der interessierten Öffentlichkeit über das Statistik-Portal[1] den Zugriff auf die Klassifikation der Wirtschaftszweige und das Güterverzeichnis für Produktionsstatistiken ermöglicht. Das mit diesem System gesammelte technische Wissen sollte genutzt werden, um einen modernen Klassifikationsserver aufzubauen. Dabei standen verschiedene Projektziele im Vordergrund:

- Das System sollte erweiterbar sein und verschiedenen Fachbereichen beziehungsweise auch anderen Organisationen den Import eigener Klassifikationen ohne aufwendige technische Einarbeitung ermöglichen.
- Das im Klassifikationsserver enthaltene Datenmaterial sollte nicht nur für Recherchezwecke zur Verfügung stehen, sondern der interessierten Öffentlichkeit auch zum Download angeboten werden können.
- Die Systemarchitektur sollte als Webanwendung mit einer modernen Drei-Schichten-Architektur skalierbar und zukunftssicher sein und aus modernen Open-Source-Technologien aufgebaut werden.
- Die erweiterbare Webservice-Schnittstelle sollte es unterschiedlichen statistischen Fachanwendungen ermöglichen, auf die gespeicherten Datenbestände zuzugreifen.
- Der bisherige Klassifikationsserver enthielt eine Thesaurus-Komponente zur Verwaltung von Wortformen und Synonymen. Dieses bislang fest in der Anwendung verankerte Modul sollte im Zuge der Neuentwicklung ausgegliedert werden, damit dessen Funktionalitäten auch

1 *www.statistik-portal.de* – Statistik-Portal der Statistischen Ämter des Bundes und der Länder.

Schaubild 1 Aufbau und Nutzung des Klassifikationsservers



anderen statistischen Fachanwendungen zur Verfügung stehen.

Das Entwicklungsprojekt wurde in verschiedene Ausbaustufen unterteilt und die Gesamtaufgabe somit in kleinere Einheiten zerlegt. Dadurch war der jeweilige Arbeitsaufwand für jede Stufe besser abzuschätzen und die Projektrisiken waren besser zu kalkulieren. Außerdem konnten durch die in den ersten Ausbaustufen gesammelten Erfahrungen zusätzliche Anforderungen an das System abgeleitet werden, die seinen Gebrauch verbesserten.

Systemarchitektur

Der Klassifikationsserver wurde als Webanwendung entwickelt, in Form einer klassischen Drei-Schichten-Architektur. Das bedeutet zum einen, dass auf die Anwendung mit einem Webbrowser zugegriffen werden kann und keine zusätzliche Software installiert werden muss. Zum anderen bietet die Architektur des Gesamtsystems ein hohes Maß an Flexibilität, da die einzelnen Komponenten für die Datenbank, die Geschäftslogik und die grafische Benutzeroberfläche voneinander getrennt und austauschbar sind. Damit kann das System insgesamt besser gewartet werden.

Die eingesetzten Programmiersprachen und Softwarekomponenten auf JAVA-Basis werden im Verbund der Statistischen Ämter des Bundes und der Länder auf breiter Fläche eingesetzt, sodass auch in Zukunft Weiterentwicklungen des Systems ohne aufwendige Einarbeitungen der beteiligten Softwareentwickler möglich sind. Die gewählte Architektur erlaubt es zudem, das System an zusätzliche Leistungsanforderungen anzupassen, indem weitere Komponenten zur Lastverteilung eingebunden werden. Einen ausführlicheren Überblick über die verwendeten Softwarekomponenten liefert ein Aufsatz von Sebastian Hilder vom Bayerischen Landesamt für Statistik und Datenverarbeitung.[2]

Der Klassifikationsserver und der Thesaurus stehen sowohl im Behördennetz als auch im Internet zur Verfügung. Die erweiterten Administrationsfunktionen im Behördennetz erlauben Fachanwendern zum Beispiel den Import zusätzlicher Klassifikationen oder die Pflege der bereits importierten Stichwörter. Neben der grafischen Benutzeroberfläche steht eine SOAP (Simple Object Access Protocol – Netzwerkprotokoll)-basierte Webservice-Schnittstelle zur Verfügung. Über diese Schnittstelle können statistische Fachanwendungen an den Klassifikationsserver gekoppelt und der Datenbestand abgefragt werden. Ein zusätzliches drittes System, der sogenannte Sync-Server, überschreibt täglich die Daten des Internet-Systems mit den aktualisierten Daten des Klassifikationsservers im Behördennetz.

Der angeschlossene Thesaurus ist für den Anwender nicht sichtbar, unterstützt ihn aber bei der Recherche in den Klassifikationen. Anfragen an den Klassifikationsserver werden intern an den Thesaurus weitergeleitet und dort ausgewertet; die Antworten werden wieder auf der Benutzeroberfläche angezeigt. Verschiedene Strategien zur Zwischenspeicherung reduzieren dabei den entstehenden Datenverkehr bei häufig wiederkehrenden Anfragen.

Datenmodell

Der Klassifikationsserver verwendet auf fachlicher Ebene das Neuchâtel-Modell für Klassifikationen[3]. Dieses Daten-

2 Hilder, S: „Deutsch-französischer Workshop über den Klassifikationsserver" in Bayern in Zahlen, Jahrgang 143, Ausgabe 11/2012, Seite 777 ff.

3 Neuchâtel Terminology Model – Classification database object types and their attributes, Version 2.1 (*www1.unece.org/stat/platform/pages/viewpage.action?pageId=14319930*, abgerufen am 15. Januar 2014).

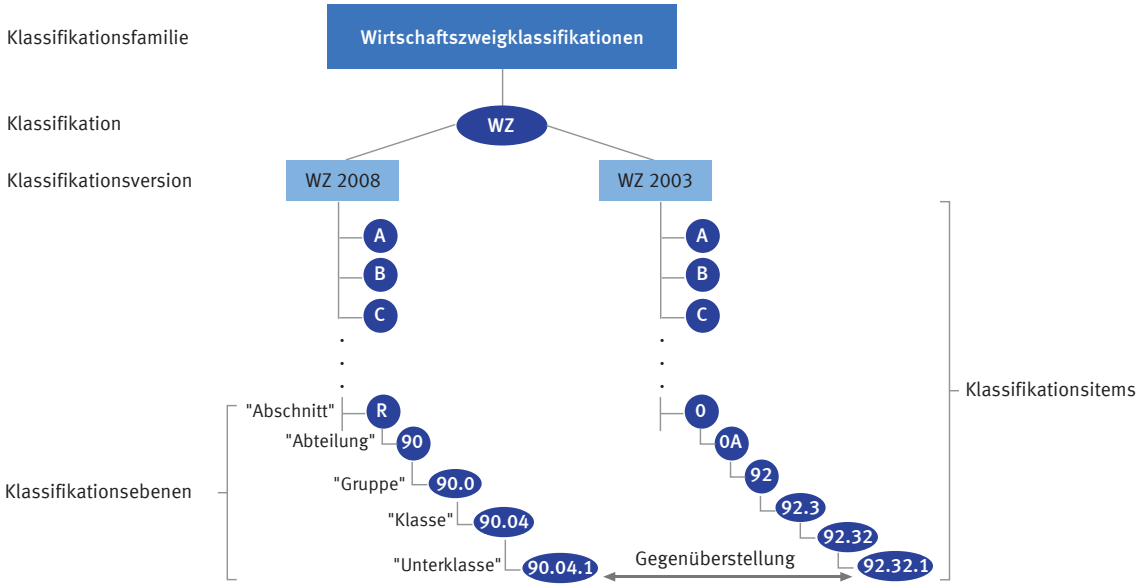
modell hat eine internationale Arbeitsgruppe entwickelt, in der auch verschiedene Statistikämter vertreten waren und die das Ziel verfolgt hat, grundlegende Konzepte in diesem Themenbereich zu definieren und die Terminologie der Klassifikationen zu vereinheitlichen.

Mit einem einzigen Datenmodell ist es einfacher, die verschiedenen Klassifikationen unterschiedlicher Fachbereiche zu verarbeiten. Die jeweiligen Objekte und ihre Eigenschaften sind klar benannt und gelten für alle Klassifikationen gleichermaßen.

Jede Klassifikation ist eingebettet in eine Klassifikationsfamilie und besteht aus verschiedenen Klassifikationsversionen. Auch Varianten von Klassifikationsversionen können gebildet werden.

Eine Klassifikationsversion ist eine Liste sich gegenseitig ausschließender Kategorien, welche die zu beobachtenden ökonomischen, sozialen oder sonstigen Tatbestände einteilen. Die Kategorien der Klassifikationsversion werden repräsentiert durch Klassifikationsitems. Jedes Item besitzt einen oder mehrere Titel, zum Beispiel einen ausführlichen Lang- und einen Kurztitel, sowie einen eindeutigen Code. Die Items sind hierarchisch in verschiedenen Ebenen angeordnet. Ihre Inhalte, aber auch die Abgrenzungen zu anderen Items, werden in den *Erläuterungen* beschrieben und können unterteilt werden in *Allgemeine Bemerkungen*, *Einschlüsse*, *Umfasst ferner* und *Ausschlüsse*. Die Bezeichnung der Felder findet sich in den zum Download angebotenen Dateien, wegen der besseren Lesbarkeit jedoch nicht in der Webansicht. Unter *Einschlüsse* wird der

Schaubild 2 Das Neuchâtel-Modell für Klassifikationen am Beispiel der Klassifikation der Wirtschaftszweige



Erläuterungen

Einschlüsse

90.04.1 Theater- und Konzertveranstalter

Diese Unterklasse umfasst:

– Tätigkeiten von Theaterveranstaltern, Varieté- und Kleinkunstveranstaltern, Konzert- und Ballettveranstaltern

Ausschlüsse

Diese Unterklasse umfasst nicht:

– Betrieb von Kinos (siehe 59.14.0)
– Künstleragenturen (siehe 74.90.0)
– Tätigkeiten von Vorverkaufsstellen (siehe 79.90.0)
– Rollenbesetzung (siehe 78.10.0)

Schlüssel	Titel
59.14.0	Kinos
74.90.0	Sonstige freiberufliche, wissenschaftliche und technische Tätigkeiten anderweitig nicht genannt
78.10.0	Vermittlung von Arbeitskräften
79.90.0	Erbringung sonstiger Reservierungsdienstleistungen

Stichwörter (Anzeige/Freitextsuche)

Stichwörter:

– Aufführung von Theaterstücken, Opern, Konzerten u.Ä. (Theater- und Konzertveranstalter)	konzert konzertveranstalter theaterstück aufführung oper theaterveranstalter
– Ballettveranstalter	ballettveranstalter
– Kleinkunstveranstalter	kleinkunstveranstalter
– Konzertveranstalter	konzertveranstalter
– Theaterveranstalter	theaterveranstalter
– Varietéveranstalter	varietéveranstalter

WZ = Klassifikation der Wirtschaftszweige.

2014 - 01 - 0054

Inhalt der Kategorie spezifiziert, während unter *Umfasst ferner* eine Liste von Grenzfällen angegeben wird, die noch zu dieser Kategorie zu zählen sind. Im Feld *Ausschlüsse* werden entsprechend Grenzfälle eingetragen, die nicht mehr in diese Kategorie fallen. Ergänzt wird diese Liste durch eine Tabelle mit Verknüpfungen zu diesen Klassifikationsitems. Das Feld *Allgemeine Bemerkungen* schließlich nimmt sämtliche Informationen zum Item auf, die zu keinem der vorherigen Felder zählen.

Neben den Klassifikationsitems mit offiziellen Codes können auch nicht offizielle Items im Datenmaterial enthalten sein, die zum Beispiel dann benötigt werden, wenn unvollständige Hierarchien mit Zwischenebenen aufgefüllt werden müssen. Diese „unechten" Items werden in der Webanwendung entweder gar nicht oder als Zwischenüberschriften ohne Code angezeigt und sind in dem Datenmaterial, das heruntergeladen werden kann, ausschließlich in der XML-Datei im CLASET-XML-Format[4] enthalten. Dieses Datenformat erlaubt die Deklaration der Itemtypen ("DummyGrouping" beziehungsweise "Uncoded"), sodass eine Verwechslung mit offiziellen Klassifikationsitems ("OfficialGrouping") verhindert wird.

Ein wichtiger Bestandteil des Datenmodells sind die Gegenüberstellungen. Eine Gegenüberstellung ist eine Menge von bidirektionalen Relationen zwischen zwei Klassifikationsitems unterschiedlicher Versionen oder Varianten. Gegenüberstellungen erlauben den direkten Vergleich zweier Versionen derselben Klassifikation und bilden somit die Entwicklung der Klassifikation im Zeitverlauf ab. Gegenüberstellungen können darüber hinaus die Beziehungen zwischen einzelnen Versionen verschiedener Klassifikationen ausdrücken und ermöglichen damit zum Beispiel den Vergleich zwischen tätigkeitsbezogenen Klassifikationen und Güterklassifikationen.

Stichwortverzeichnis

Zusätzlich zu den Erläuterungen können jedem Klassifikationsitem beliebig viele Stichwörter zugeordnet werden, die bei der Einordnung statistischer Tatbestände in die Klassifikationsversion unterstützen. Während Klassifikationsversionen unter Umständen mehrere Jahre Gültigkeit besitzen, können die Stichwortverzeichnisse im Laufe der Zeit an die Realität angepasst werden.

Neben den auf der Benutzeroberfläche des Klassifikationsservers angezeigten Stichwörtern ist eine zusätzliche Liste von Stichwörtern hinterlegt, die separat gepflegt wird. Diese Stichwörter werden ausschließlich für die Freitextsuche verwendet und über den angeschlossenen Thesaurus zusätzlich einer Grundformreduktion unterzogen. Obwohl es sich technisch um zwei getrennte Stichwortlisten handelt, sind die beiden Listen inhaltlich nahezu identisch. Die Unterschiede betreffen unter anderem die in den Stichwörtern beschriebenen Ausschlüsse. So wird beispielsweise das Stichwort „Anlasser von Verbrennungsmotoren (nicht für Kraftfahrzeuge), Handelsvermittlung" des WZ2008-Items 46.14.1 zu „Anlasser von Verbrennungsmotoren, Handelsvermittlung" reduziert. Ohne die manuelle Anpassung des Stichworteintrages würde bei der Suche nach „Kraftfahrzeug" dieser Begriff für die Ergebnismenge mitberücksichtigt, obwohl Kraftfahrzeuge explizit ausgeschlossen wurden.

Um das Ergebnis der Freitextsuche weiter zu verbessern, werden die Stichwörter in deutscher Sprache einer automatisierten Grundformreduktion unterzogen, wobei sogenannte Stoppwörter entfernt und beispielsweise Substantive in den Nominativ Singular überführt werden. Das oben betrachtete Stichwort wird durch die Grundformreduktion schließlich zu „anlasser verbrennungsmotor handelsvermittlung".

Funktionsumfang

Mit der Neuentwicklung des Klassifikationsservers wurde der Funktionsumfang stark ausgebaut. Die neue Anwendung erweitert die Liste der Standardanwendungen im Verbund der Statistischen Ämter des Bundes und der Länder und leistet perspektivisch als zentraler Datenspeicher nationaler Standardklassifikationen einen wichtigen Beitrag zur Optimierung der statistischen Produktionsprozesse.

Mandantenfähigkeit

Der Klassifikationsserver ist ein mandantenfähiges System. Das bedeutet, dass die Datenbestände der verschiedenen Mandanten, im Klassifikationsserver Einrichtungen genannt, vollständig voneinander getrennt bearbeitet und verwaltet sowie die Zugriffsberechtigungen eigenständig festgelegt werden können. Über ein integriertes Anwendungsprotokoll können vom Systemadministrator nachträglich die Bearbeitungen der Datenbestände nachvollzogen werden. Dem Anwender erschließen sich die unterschiedlichen Zuständigkeiten der verschiedenen Klassifikationen durch die Metainformationen auf der Auswahlseite. Die Darstellung der Inhalte ist durch das verwendete Datenmodell für alle Klassifikationen einheitlich.

Neue Importfunktion

Für den Klassifikationsserver wurde eine Schnittstelle entwickelt, die es den Fachbereichen ermöglicht, neue Klassifikationen in das System zu importieren. Die für die Aufbereitung des Datenmaterials benötigte Office-Anwendung steht dabei auf jedem Standardarbeitsplatzrechner zur Verfügung. Als Importformat dient das CLASET-XML-Format, das als technische Implementierung des fachlichen Neuchâtel-Modells unter anderem von Ramon[5], dem Klassifikationsserver von Eurostat, verwendet wird. Die mit der Office-Anwendung erstellten Eingabedateien werden über ein Skript in dieses Format umgewandelt und können anschließend in das System importiert werden.

Das CLASET-XML-Format ist mehrsprachig ausgelegt und auch die Schnittstelle des Klassifikationsservers ermöglicht den Import von Daten in beliebiger Sprache. Somit können im Prinzip auch internationale Klassifikationen impor-

4 CLASET-Formatbeschreibung des Statistischen Amtes der Europäischen Union (Eurostat): *http://ec.europa.eu/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=DSP_CLASET_PAGE* (abgerufen am 15. Januar 2014).

5 *http://ec.europa.eu/eurostat/ramon/index.cfm*, abgerufen am 16. Januar 2014.

tiert werden. Von der Möglichkeit wird derzeit jedoch kein Gebrauch gemacht, da die pflegenden Einrichtungen nicht in allen Fällen über Neuerungen in den Versionen informieren und somit die Aktualität der Daten im Klassifikationsserver nicht gewährleistet werden könnte. Stattdessen wird in den Metainformationen auf der Auswahlseite der jeweiligen Klassifikationsversion oder -variante auf die gegebenenfalls vorhandenen internationalen Referenzklassifikationen verlinkt.

Historisierung der Daten

Der Datenbestand des Klassifikationsservers wird historisiert, sodass Änderungen an den Inhalten einer Klassifikationsversion sowie den zugehörigen Stichwörtern nachvollziehbar bleiben. Während über die Webanwendung die jeweils aktuellsten Daten angezeigt werden, können über die Webservice-Schnittstelle zusätzlich auch Daten abgerufen werden, die zu einem früheren Zeitpunkt gültig waren. Da sich Stichwörter häufiger ändern können als die zugrunde liegenden Klassifikationsitems, werden auf der Auswahlseite jeder Klassifikationsversion die Zeitpunkte der Änderungen getrennt ausgewiesen.

Neue Benutzeroberfläche

Die Recherchefunktionen wurden im neuen Klassifikationsserver weiterentwickelt. Bereits auf der Auswahlseite jeder Klassifikationsversion werden zusätzliche Metadaten wie Eigentümer und Rechtsgrundlagen sowie Angaben zur letzten Aktualisierung übersichtlich dargestellt. Ebenso finden sich Informationen zu den Lizenzbestimmungen und ausführliche Vorbemerkungen. Die Benutzeroberfläche steht sowohl in deutscher als auch in englischer Sprache zur Verfügung und die kontextsensitive Onlinehilfe sowie die ausführliche Dokumentation unterstützen den Anwender bei der Nutzung des Systems. Über ein Formular können die Anwender außerdem Kontakt zu dem jeweils zuständigen Fachbereich aufnehmen.

Die Webseiten des Klassifikationsservers verfügen über sogenannte Direktlinks, sodass die jeweilige Seite eines ausgewählten Klassifikationsitems zitiert und auf die zugehörige Internetadresse verwiesen werden kann. Die Inhalte des Klassifikationsservers sind dadurch auch über Suchmaschinen im Internet auffindbar.

Wird eine neue Klassifikationsversion importiert oder eine bereits vorhandene verändert, kann der Fachbereich die Änderungen textlich beschreiben und in der Webanwendung hinterlegen. Zusätzlich können die Anwender über einen RSS-Feed in Kurzform über die Änderungen benachrichtigt werden. Die Verknüpfung zum Abonnieren des Feeds wird auf der Auswahlseite der Klassifikation dargestellt.

Erweitertes Downloadangebot

Die Exportdateien erlauben dem Anwender die Analyse beziehungsweise Weiterverarbeitung des Datenmaterials außerhalb des Klassifikationsservers; sie werden nach dem

Schaubild 3 Kompakte Darstellung eines Klassifikationsitems, das über die Gliederung ausgewählt wurde

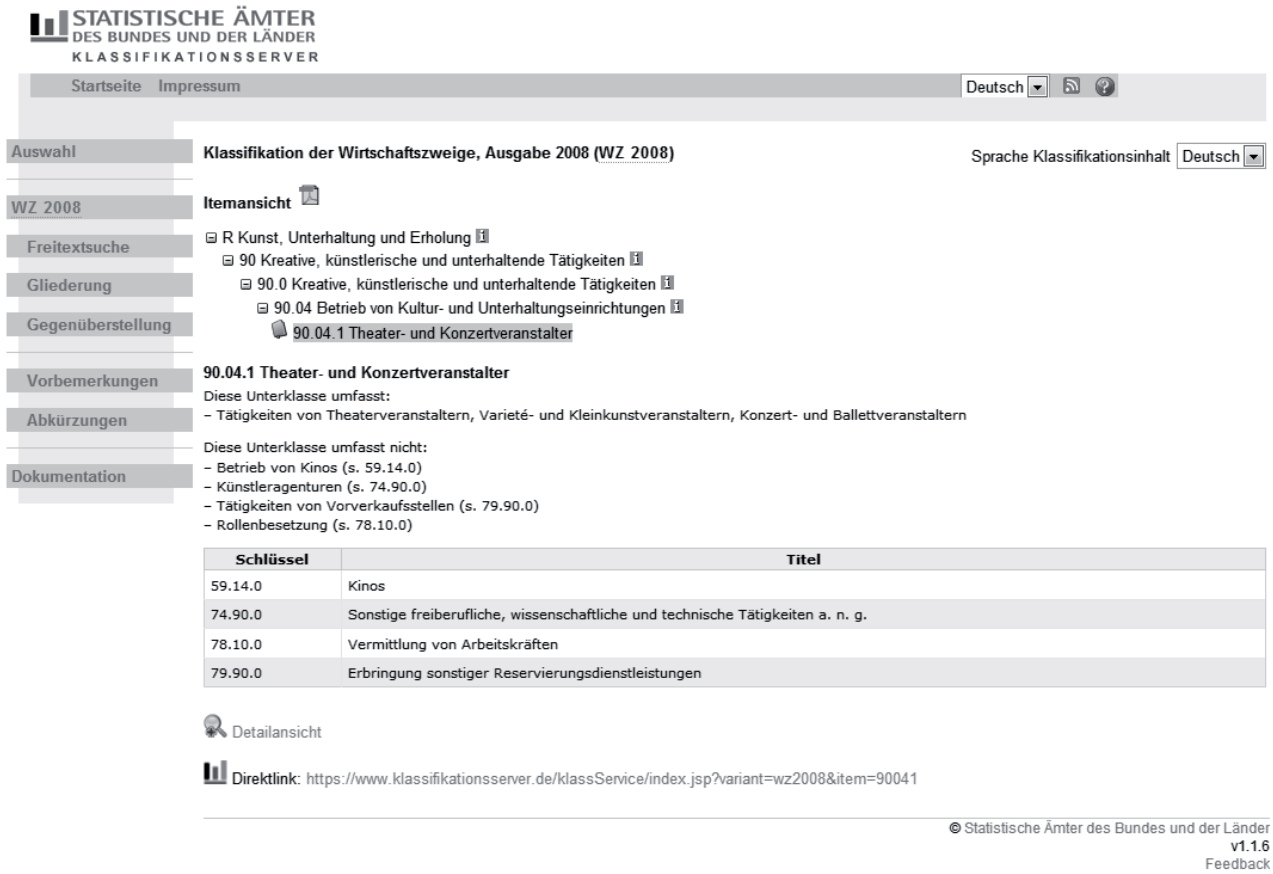
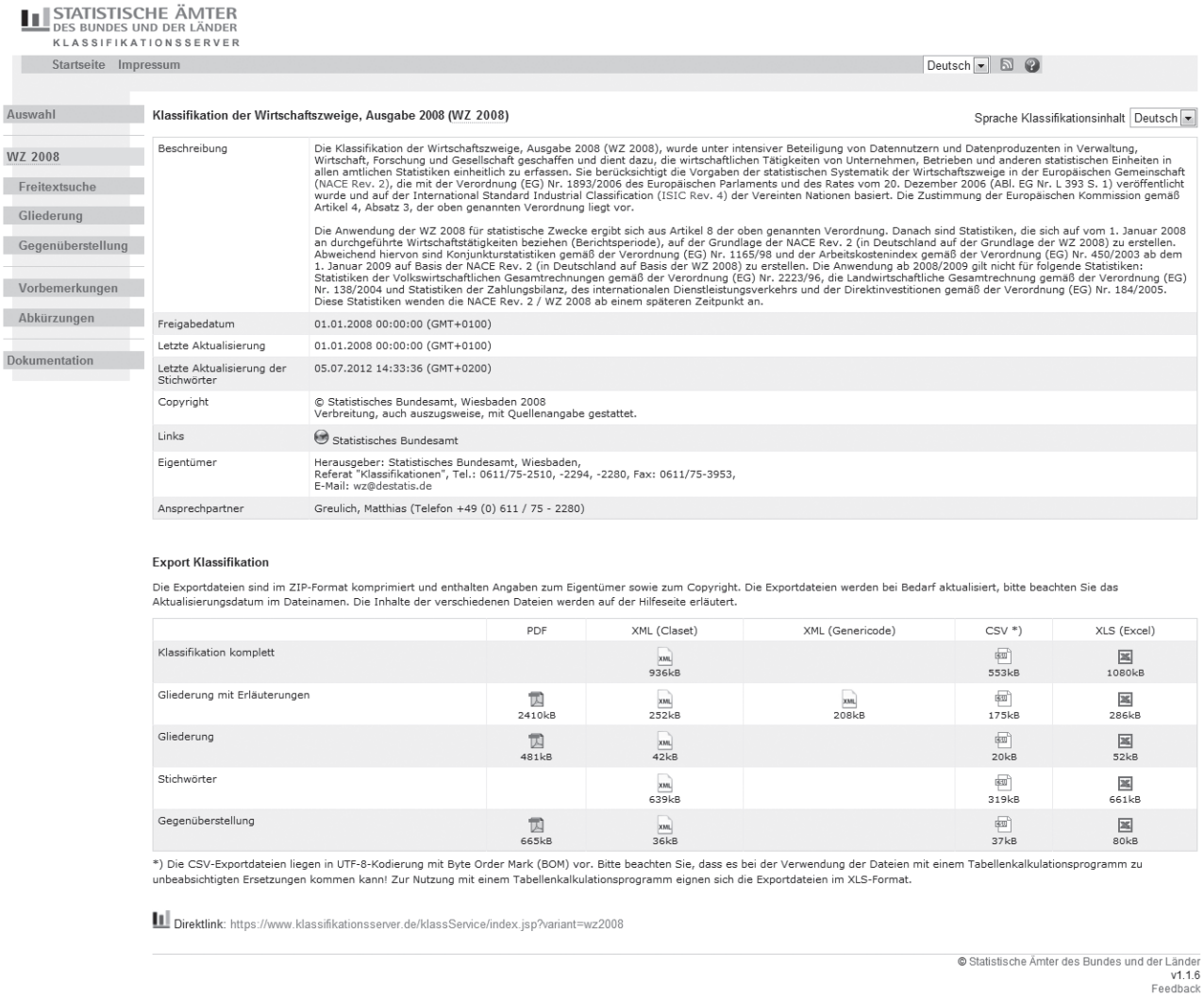


Schaubild 4 Bildschirmseite mit den Metainformationen und Exportdateien der Klassifikationsversion WZ 2008 sowie den verfügbaren Funktionen in der Menüleiste auf der linken Seite



Import der Daten beziehungsweise nach Änderungen automatisch erstellt. Neben zwei verschiedenen XML-Datenformaten kann auch das textbasierte CSV-Format gewählt werden. Außerdem steht eine Exportdatei zur Verfügung, die mit gängigen Tabellenkalkulationsprogrammen verarbeitet werden kann. Alle Dateien werden in komprimierter Form bereitgestellt, um die Downloadzeiten zu verkürzen. Das XML-Exportformat "OASIS Genericode" ermöglicht es den Fachbereichen, die importierten Klassifikationsdaten in das XRepository der Bundesstelle für Informationstechnik des Bundesverwaltungsamtes bereitzustellen. So steht zum Beispiel bereits die Klassifikation der Wirtschaftszweige in den Versionen WZ 2003 und WZ 2008 im XRepository[6] zur Verfügung.

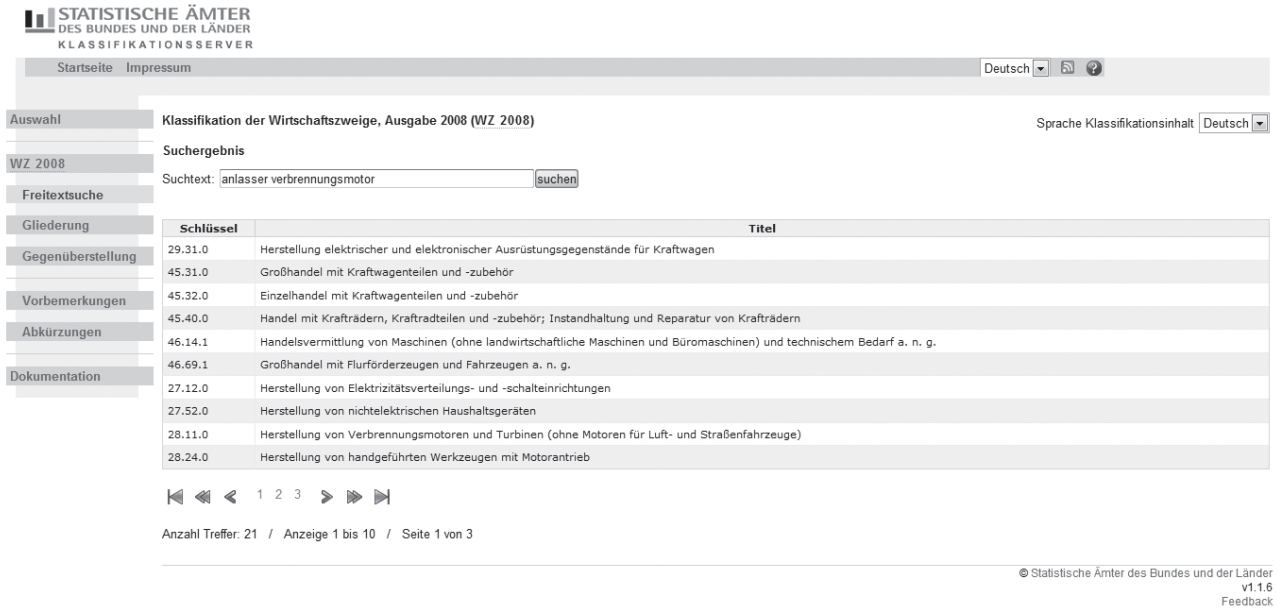
Suchfunktion

Die Freitextsuche erlaubt über die zu jedem Klassifikationsitem hinterlegten Stichwörter die komfortable Recherche nach bestimmten Items einer Klassifikationsversion. Diese Funktion steht als Webservice-Methode auch anderen Fachanwendungen zur Verfügung. In der Suchmaske können beliebig viele Suchbegriffe eingegeben werden. Je mehr Begriffe eingegeben werden, desto präziser kann der Suchalgorithmus nach passenden Items suchen. Die Suchbegriffe werden zunächst in ihre Grundform transformiert und dann mit den hinterlegten Stichwörtern verglichen, wobei für jedes Item ein Ähnlichkeitswert berechnet wird. Je mehr Suchbegriffe mit den Stichwörtern eines Items übereinstimmen, desto höher wird dieses Item in der Ergebnisliste, die bis zu 100 Einträge umfassen kann, angezeigt. Liefert eine Suchanfrage kein Ergebnis zurück, wird die Suche automatisch mit den Synonymen der Suchbegriffe wiederholt. Die Synonyme sind im nachgelagerten Thesaurus hinterlegt und können von den Fachbereichen gepflegt werden.

Das Benutzerkonzept des Thesaurus erlaubt dabei die strikte Trennung der enthaltenen Sprachbestände. Da jedoch Thesaurus und Klassifikationsserver separate Systeme sind, besteht im Bedarfsfall für verschiedene Klassifikationsser-

6 Klassifikation der Wirtschaftszweige im XRepository: *www.xrepository.deutschland-online.de/Inhalt/urn:uuid:f9c22524-5516-4c61-92af-a969722cd687.xhtml*, abgerufen am 16. Januar 2014.

Schaubild 5 Mit der Freitextsuche können Klassifikationsitems anhand von Suchbegriffen ermittelt werden



ver-Mandanten die Möglichkeit, Thesaurus-Sprachbestände gemeinsam zu nutzen. Dies ist zum Beispiel eine sinnvolle Vorgehensweise bei verwandten Klassifikationen.

Gegenüberstellungen

Der Fachbereich kann Gegenüberstellungen zu Klassifikationsitems früherer Versionen oder zu anderen Klassifikationen definieren. Beim Import der Daten werden in der Datenbank logische Verknüpfungen zwischen den Klassifikationsitems hergestellt. Damit ist gewährleistet, dass die referenzierten Objekte in der Datenbank auch tatsächlich existieren.

Der Anwender der Webanwendung kann entscheiden, ob er sich die vollständige Liste der Gegenüberstellungen anzeigen lassen möchte oder ausschließlich beispielsweise die früheren Versionen des ausgewählten Klassifikationsitems. Die einzelnen Einträge können angeklickt werden, sodass direkt zum gegenübergestellten Item gesprungen werden kann.

Erfahrungen

Bereits während der Entwicklungsphase wurde intensiv daran gearbeitet, die in elektronischer Form vorliegenden Klassifikationen an das neue Datenmodell anzupassen, um sie im Klassifikationsserver bereitstellen zu können. Da bislang kein einheitliches Datenmodell zum Einsatz kam, war der Aufwand – bedingt durch die Größe der Klassifikationen – verhältnismäßig hoch. Eine wesentliche Herausforderung bestand darin, dass das Datenmaterial ursprünglich speziell für die Erstellung der gedruckten Ausgabe der Klassifikationen aufbereitet wurde. Dabei wurden unter anderem zur besseren Lesbarkeit einzelne Ebenen des Datenmaterials entfernt, die für den Import in den Klassifikationsserver nachträglich wieder hinzugefügt werden mussten.

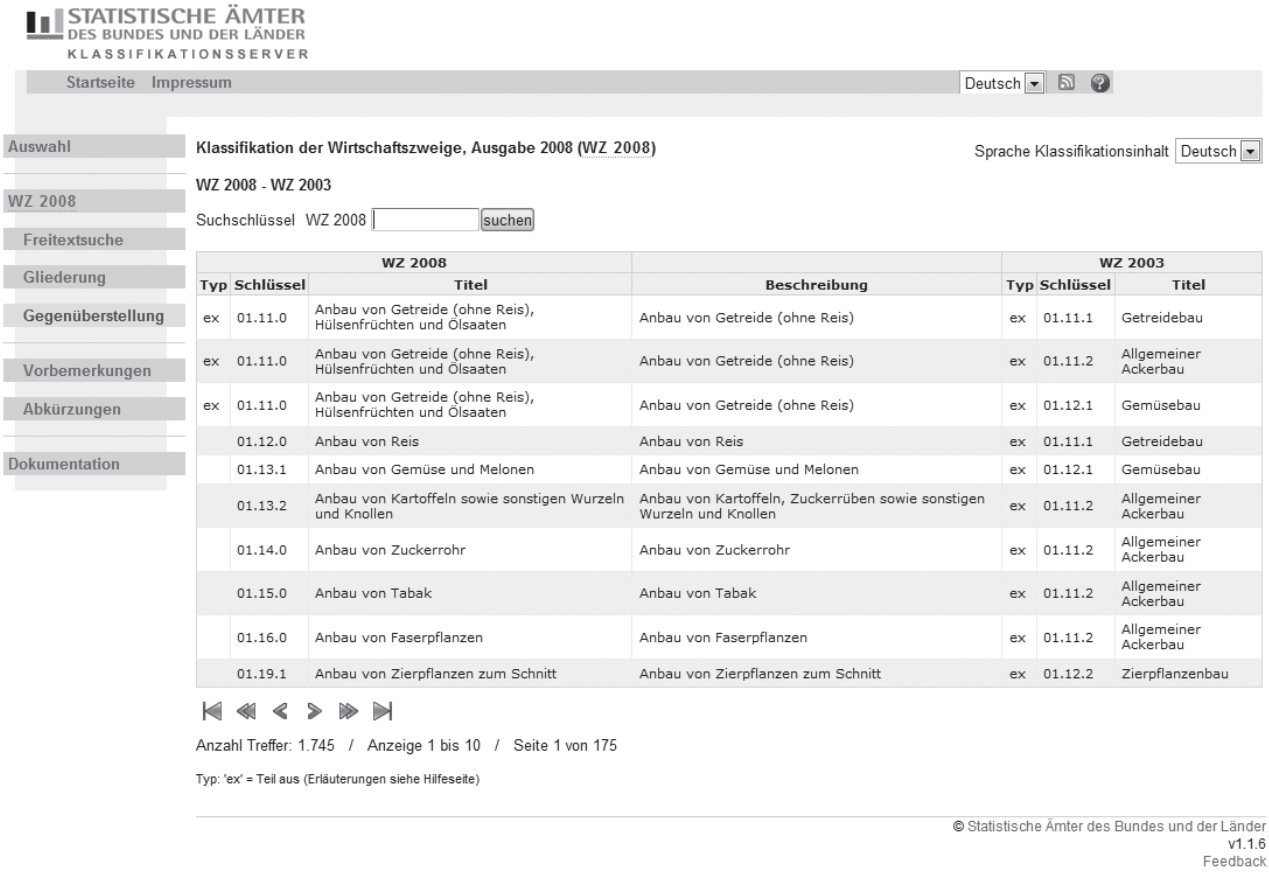
Eine eigene Clientanwendung zur Pflege von Klassifikationsdaten könnte diese Arbeiten wirkungsvoll unterstützen, zum Beispiel durch die Definition von Bildungsregeln für Itemcodes und die kontinuierliche Prüfung der Hierarchien während der Bearbeitung.

Es hat sich gezeigt, dass gerade die umfangreiche Klassifikation der Wirtschaftszweige als Pilotanwendung für den Klassifikationsserver gut geeignet war. So konnten wertvolle Erkenntnisse gewonnen werden, die zur Verfeinerung der Importschnittstelle und zu einer besseren Benutzbarkeit des Systems führten. Die gesammelten Erfahrungen des Fachbereichs wurden dokumentiert und stehen anderen Fachbereichen als Hilfestellung bei der Anpassung des eigenen Datenmaterials zur Verfügung.

Die Arbeit an der Klassifikation der Wirtschaftszweige mit ihrem umfangreichen Stichwortverzeichnis hat während der Entwicklungsphase auch offenbart, dass die Suchfunktion nicht in allen Fällen optimale Suchergebnisse lieferte. So waren einerseits die Stichwörter noch nicht bestmöglich für den Klassifikationsserver angepasst (siehe Abschnitt Stichwortverzeichnis) und andererseits der Algorithmus zur Ermittlung der Grundform eines Stichworts noch nicht ideal ausgestaltet. Zusätzliche Bemühungen in beide Richtungen führten zu verbesserten Suchergebnissen. Obwohl der zugrunde liegende Algorithmus bereits im Vorgängersystem vorhanden war, sorgte die Suche mit dem berechneten Ähnlichkeitswert insgesamt für Irritationen, sodass weitere Verbesserungen der Funktion geplant sind. Trotzdem wird man eine optimale Lösung nie erreichen, sondern sich dieser bestenfalls durch manuelle Anpassung der Stichwörter annähern können.

Der in der Pilotphase betriebene Aufwand bei der Anpassung des Datenmaterials an den Klassifikationsserver wird bei kleineren Klassifikationen weitaus geringer ausfallen. Hinzu kommt, dass gerade bei unregelmäßig erscheinenden Klassifikationen der zeitliche Druck begrenzt ist und die

Schaubild 6 Die Anzeige der Gegenüberstellungen erlaubt den direkten Vergleich zweier Klassifikationsversionen



WZ 2008				WZ 2003		
Typ	Schlüssel	Titel	Beschreibung	Typ	Schlüssel	Titel
ex	01.11.0	Anbau von Getreide (ohne Reis), Hülsenfrüchten und Ölsaaten	Anbau von Getreide (ohne Reis)	ex	01.11.1	Getreidebau
ex	01.11.0	Anbau von Getreide (ohne Reis), Hülsenfrüchten und Ölsaaten	Anbau von Getreide (ohne Reis)	ex	01.11.2	Allgemeiner Ackerbau
ex	01.11.0	Anbau von Getreide (ohne Reis), Hülsenfrüchten und Ölsaaten	Anbau von Getreide (ohne Reis)	ex	01.12.1	Gemüsebau
	01.12.0	Anbau von Reis	Anbau von Reis	ex	01.11.1	Getreidebau
	01.13.1	Anbau von Gemüse und Melonen	Anbau von Gemüse und Melonen	ex	01.12.1	Gemüsebau
	01.13.2	Anbau von Kartoffeln sowie sonstigen Wurzeln und Knollen	Anbau von Kartoffeln, Zuckerrüben sowie sonstigen Wurzeln und Knollen	ex	01.11.2	Allgemeiner Ackerbau
	01.14.0	Anbau von Zuckerrohr	Anbau von Zuckerrohr	ex	01.11.2	Allgemeiner Ackerbau
	01.15.0	Anbau von Tabak	Anbau von Tabak	ex	01.11.2	Allgemeiner Ackerbau
	01.16.0	Anbau von Faserpflanzen	Anbau von Faserpflanzen	ex	01.11.2	Allgemeiner Ackerbau
	01.19.1	Anbau von Zierpflanzen zum Schnitt	Anbau von Zierpflanzen zum Schnitt	ex	01.12.2	Zierpflanzenbau

Schnittstellen inzwischen einen gewissen Reifegrad erreicht haben und ausführlich dokumentiert sind.

Der unmittelbare Nutzen, den die Bereitstellung von Klassifikationen im Klassifikationsserver für die amtliche Statistik bietet, stellt sich spätestens bei der Neu- oder Weiterentwicklung von Fachanwendungen ein, die intern diese Daten nutzen. Das Datenmodell erlaubt es, vereinheitlichte und wiederverwendbare Verarbeitungsroutinen zu entwickeln. Erste Fachanwendungen wurden bereits erfolgreich über die Webservice-Schnittstelle angebunden.

Außerdem wird die Außendarstellung verbessert, da in der Öffentlichkeit ein wachsendes Interesse an maschinenlesbaren statistischen Daten und Metadaten zu beobachten ist.

Ausblick

Mit der Freischaltung des Klassifikationsservers im Behördennetz sowie im Internet wurden die Grundlagen geschaffen, eine zentrale Plattform für nationale Standardklassifikationen bereitzustellen. In Zusammenarbeit mit der Bundesagentur für Arbeit wurde mit der Klassifikation der Berufe inzwischen eine weitere Standardklassifikation bereitgestellt. Auch das Güterverzeichnis für Produktionsstatistiken ist mittlerweile im Klassifikationsserver verfügbar.

Die Entwicklung des Klassifikationsservers wurde international mit großem Interesse verfolgt. So zeigte sich insbesondere das französische Statistikamt INSEE an einer Zusammenarbeit interessiert. In einem gemeinsamen Workshop mit dem Entwicklerteam vom Bayerischen Landesamt für Statistik und Datenverarbeitung wurden die technischen Aspekte des Klassifikationsservers ausführlich diskutiert.[7] Auch dem Australian Bureau of Statistics wurde der Klassifikationsserver vorgestellt. Dort steht insbesondere das Neuchâtel-Modell als fachliche Grundlage im Fokus des Interesses.

Geplante Weiterentwicklungen zielen insbesondere darauf ab, die Standardklassifikationen auch im Datenerhebungsprozess bereitzustellen. Weiterhin soll die Webservice-Schnittstelle erweitert werden, damit statistische Fachanwendungen noch besser unterstützt werden können.

Das Potenzial des Klassifikationsservers kann sich nur dann voll entfalten, wenn es gelingt, weitere Klassifikationen in das System zu integrieren – mit möglichst wenig Zusatzaufwand für die beteiligten Fachbereiche. Die größte technische Herausforderung stellt bislang die fehlende Unterstützung des Klassifikationsservers bei der Erstellung von Druckerzeugnissen dar. Auch in Zukunft bleibt die gedruckte Fassung der Klassifikationen ein wichtiges, oft sogar primäres Produkt. Aus Sicht der Fachbereiche kann der Klassifikationsserver insbesondere dann sinnvoll eingesetzt werden, wenn der gesamte Lebenszyklus einer Klassifikationsver-

7 Siehe Hilder, S. (Fußnote 2).

sion von der Erstellung bis zur Veröffentlichung unterstützt wird.

Die Erzeugung einer Druckvorlage ist zum Beispiel bei der jährlichen Aktualisierung einer Klassifikationsversion ein aufwendiger und auch zeitkritischer Prozess. Die vielfältigen Gestaltungsregeln und teilweise von Klassifikation zu Klassifikation unterschiedlichen Darstellungen stehen bislang einer vollautomatischen Lösung im Wege. Eine Annäherung an dieses Problem und die Implementierung einer semiautomatischen Lösung könnte bereits zu einer deutlichen Erleichterung beitragen und zu einer stärkeren Akzeptanz des Systems führen.

Im Hinblick auf die Entwicklungen im Bereich des Semantischen Webs ist auch eine Erweiterung des Downloadangebotes und Unterstützung des Resource Description Frameworks (RDF)[8] denkbar, um die maschinelle Verarbeitung der Daten weiter zu verbessern. Auch die kontinuierliche Weiterentwicklung des in der amtlichen Statistik verbreiteten Standards SDMX[9] könnte künftig Auswirkungen auf die Gestaltung des Klassifikationsservers haben.

Insgesamt ist ein großes Interesse an statistischen Metadaten in maschinenlesbarer Form zu beobachten. Mit dem Klassifikationsserver steht ein Werkzeug zur Verfügung, das nicht nur die Recherche in Standardklassifikationen ermöglicht, sondern diese Daten auch verschiedenen Zielgruppen – von interessierten Bürgerinnen und Bürgern bis hin zu Herstellern behördenspezifischer Software – in maschinenlesbarer Form zur Verfügung stellt.

8 Resource Description Framework: *https://de.wikipedia.org/wiki/Resource_Description_Framework*, abgerufen am 16. Januar 2014.

9 SDMX – Statistical Data and Metadata Exchange: *www.sdmx.org*, abgerufen am 16. Januar 2014.

Auszug aus Wirtschaft und Statistik

Herausgeber
Statistisches Bundesamt, Wiesbaden

www.destatis.de

Schriftleitung
Dieter Sarreither,
Vizepräsident des Statistischen Bundesamtes

Redaktion: Ellen Römer
Telefon: +49 (0) 611 / 75 23 41

Ihr Kontakt zu uns
www.destatis.de/kontakt

Statistischer Informationsservice
Telefon: +49 (0) 611 / 75 24 05

Abkürzungen

WiSta	=	Wirtschaft und Statistik
MD	=	Monatsdurchschnitt
VjD	=	Vierteljahresdurchschnitt
HjD	=	Halbjahresdurchschnitt
JD	=	Jahresdurchschnitt
D	=	Durchschnitt (bei nicht addierfähigen Größen)
Vj	=	Vierteljahr
Hj	=	Halbjahr
a. n. g.	=	anderweitig nicht genannt
o. a. S.	=	ohne ausgeprägten Schwerpunkt
St	=	Stück
Mill.	=	Million
Mrd.	=	Milliarde

Zeichenerklärung

p	=	vorläufige Zahl
r	=	berichtigte Zahl
s	=	geschätzte Zahl
–	=	nichts vorhanden
0	=	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
.	=	Zahlenwert unbekannt oder geheim zu halten
. . .	=	Angabe fällt später an
X	=	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
I oder —	=	grundsätzliche Änderung innerhalb einer Reihe, die den zeitlichen Vergleich beeinträchtigt
/	=	keine Angaben, da Zahlenwert nicht sicher genug
( )	=	Aussagewert eingeschränkt, da der Zahlenwert statistisch relativ unsicher ist

Abweichungen in den Summen ergeben sich durch Runden der Zahlen.

© Statistisches Bundesamt, Wiesbaden 2014

Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.